DOCUMENT RESUME

CG 003 725

ED 028 468

By-Nyberg, V. R.

The Reliability of Essay Grading.

Canadian Council for Research in Education, Ottawa (Ontario).
Pub Date Jun 68
Note-6p.; Paper presented at the Sixth Canadian Conference on Educational Research, Ste. Foy, Quebec, June 1968.
EDRS Price MF-$0.25 HC-$0.40
Descriptors-Analysis of Variance, *Essay Tests, *Evaluation Criteria, Factor Analysis, Grading, *Reliability, Research, *Test Construction, Testing, Test Interpretation, Test Results, Tests, *Test Validity

The following problems in the field of essay grading have persisted: (1) low reliability of grading, (2) deciding upon elements or variables to consider, and (3) deciding on the weight to assign to each variable. Described are two aspects of a study on essay grading at the high school level in Alberta, Canada: (1) the reliability of scoring procedures, and (2) the effectiveness of the procedures with respect to educational objectives. The grade 12 essay examination was written by 13,000 students. It was scored by 48 readers according to a fixed pattern employing 22 variables grouped according to grammar and content. The following statistical procedures were performed: (1) correlations of scores given by each reader, (2) a factor analysis on the variables to determine what underlying elements were present, (3) estimations of reliability of scoring by use of correlation means, and (4) an analysis of variance on scores. The effectiveness of the variables was judged. Variables were then grouped according to six factors, and tables to factor loadings and factor descriptions were developed. Style-content variables were found to be very heavily weighted and as this was against the intention of the author, another study, employing a different weighting technique is planned. (JS)

THE RELIABILITY OF ESSAY GRADING*

Despite efforts of researchers, three problems in the field of essay grading have persisted. These problems are:

1) low reliability of grading

2) deciding upon the elements or variables to consider while scoring

3) deciding on what weight should be given to each of the variables

Earlier studies such as those reported by Starch and Elliot (1912), Darsie (1922), and Hulten (1925) were concerned almost entirely with low reliability of grading. Later studies dealt with one aspect of the problem of what variables should be employed in grading essays, and, in particular, whether "wholistic" or "atomistic" approaches should be used. The underlying purpose in such studies, however, was usually to find a way of improving reliability of grading. Studies in this area were reported by Cast (1939; 1940), Morrison and Vernon (1941), Coward (1952), Torgerson and Green (1952), Huddleston (1954), Remondino (1959), Diederich, French and Carleton (1961), and Coffman and Kurfman (1968).

The purpose of this paper is to describe two aspects of a study conducted in the broad area of essay grading at the high school leaving level in the Province of Alberta. One aspect is the reliability of the scoring procedures, the other is the effectiveness of the procedures with respect to the educational objectives. The branches of the various provincial departments of education charged with the task of grading examinations are well aware of the problem of unreliability of scoring, and most of them have devised special procedures for improving the reliability. The study reported here might therefore be of interest to a number of education departments.

The Department of Education in Alberta administers a battery of achievement examinations for Grade XII students. Among the examinations is a two-hour test that involves writing an original essay. The essays are scored by a committee of Grade XII English teachers selected by the department. The study reported here centered about the scoring of these essays.

In June of 1964 approximately 13,000 students wrote the Grade XII essay examination. Students were given two topics, one of which was to be developed in an essay of 300 words or less. The scoring was accomplished by a group of 48 readers, divided into two groups. One group was responsible for grading mechanics of English, the other for grading style and content. Each essay was read twice, but at each reading it was scored for different things.

Special procedures have been adopted in an effort to improve the reliability of scoring. Minimum qualifications of training and teaching experience for readers have been specified, but, more important, the scoring is done according to a fixed pattern employing a total of 22 variables. The essays are identified by numbers, only, during the scoring operation, and the readers work under the supervision of two chairmen, one in charge of the 'mechanics' group, and the other in

---

*CCRE is pleased to bring you this paper. The ideas expressed are those of the author.

charge of the 'style-content' group. The chief responsibility of the chairmen is to see that the standard of scoring is as uniform as possible. Before the scoring actually begins approximately one half day is spent in discussion and definition of the 22 variables.

The procedure adopted in the study was to select a sample of 103 essays, all on one topic, and to have each of the readers score all of them independently. It was then possible to correlate scores of readers, and also to calculate correlations among the 22 variables. A factor analysis, based on the matrix of correlations between pairs of variables, was carried out for the purpose of determining what underlying variables were present. It was possible, also, to determine how much each variable contributed to the total score and to see whether the stated goal of having half the score being contributed by 'mechanics' and the other half from 'style-content' had been achieved.

An estimate of the reliability of the scoring was determined from the matrix of correlations between pairs of readers, by finding the mean of these correlations. For the 'mechanics' group the reliability was .77, and for the 'style-content' group it was .60. A suggestion at this point that the marking of the 'mechanics' group was more satisfactory becomes untenable when one examines the means and standard deviations. Since all the readers scored the same 103 essays the mean scores for the 'mechanics' readers should be equal, as should the mean scores for the 'style-content' group. Differences in means therefore reflect differences in standards among markers. For the 'mechanics' group the means varied from 40.0 to 72.4, and the standard deviations from 19.8 to 28.7. For the 'style-content' group the means ranged from 63.6 to 72.4 and the standard deviations from 9.3 to 13.1. This would indicate that certain exacting readers consistently found almost twice as many mechanics errors as certain lenient readers.

An analysis of variance procedure, using one of Winer's (1962) models, shows that the variation in marks from one essay to another accounted for 59% of the total variation. The remaining 41%, however, was not attributable to differences among readers.

In all, each essay was graded with respect to 22 variables. The effectiveness of these variables was judged first through use of a factor analytic procedure. Interpretation of the factors was attempted on the basis of a varimax rotation, with loadings in excess of .65, only, being considered. The factor loadings are as follows:

TABLE I: FACTOR LOADINGS OF GRADING VARIABLES

	Variable	Factors I	II	III	IV	V	VI
↑ Mechanics →	Spelling						.67
	Punctuation					.84	
	Word Usage		.86				
	Grammar		.76				
	Sentence Errors		.76				
	Form				.9		
↓ Mechanics ←	Significance	.94					
	Relevance	.90					
	Originality	.89					

(cont'd. on next page)

	Variable	Factors I	II	III	IV	V	VI
Style-content	Plan			.95			
	Relation of plan to essay			.93			
	Introduction	.84					
	Order	.92					
	Emphasis	.95					
	Conclusion	.86					
	Vividness of words	.94					
	Figures of speech	.89					
	Vocabulary	.93					
	Sentence structure	.93					
	Sentence beginnings	.89					
	Economy	.91					
	Total impression	.96					

Owing to the number of comparatively high loadings it was not too difficult to find reasonable labels for the factors.

Factor I was called 'general proficiency in style and content' or 'general impression of style and content'. The variable 'general impression' had a very high loading, but it was not clear whether this score was awarded on the basis of an unconscious totalling of the subscores, or whether, in reading an essay, a marker quickly formed a general impression, then made the subscores fit his impression.

Factor II was called 'higher mechanical skills in writing'.

Factor III was obviously the essay plan. It was noted that this factor seemed unrelated to other variables. For the top half of the essay scores the variable 'plan' was generally negatively correlated with other variables. One possible explanation might be that weak students were drilled so as to complete the plan after the essay was completed. It seemed a reasonable conclusion that the two variables that made up Factor III be redefined, discarded, or measured in another way.

Factors IV to VI, respectively, contained only one variable each, and were labelled, respectively; form, punctuation, and spelling. The variable labelled 'form' tended to be negatively correlated with other scores, therefore it seemed reasonable to recommend dropping this variable from the list used in grading essays.

A study of the contribution of each essay variable to the total score was very revealing. It was mentioned earlier in this paper that a decision had been made by the curriculum makers that the 'mechanics' and 'style-content' scores should be equally weighted. The implementation of this decision had been rather naive. The highest possible raw score for the two sections were made equal, on the assumption that this would ensure equal weighting. Furthermore, raw scores were assigned to each of the variables on the assumption that each of these would be weighted in proportion to the value assigned.

Calculations were made so that comparisons could be made between expected and actual contributions of each of the variables to the total score. The comparisons are contained in the following figures:

TABLE II: STATISTICS RELATED TO ESSAY SCORING VARIABLES

VARIABLES	Max. Score Possible	Mean Score Awarded	S.D. of Scores	Contribution to Total Score (%) Expected	Actual
Mechanics					
Spelling	39	20.9	10.2	11.1	27.6
Punctuation	26	15.7	3.3	7.4	8.5
Word Usage	24	16.3	2.9	6.9	7.7
Grammar	33	24.2	3.9	9.5	12.2
Sentence Errors	39	29.4	4.5	11.1	14.5
Form	14	9.7	2.2	4.0	3.3
Total Mechanics	175*	116.2		50.0	73.8
Style-Content					
Significance	10	6.0	.6	4.0	1.6
Relevance	10	6.3	.6	4.0	1.6
Originality	5	1.8	.7	2.0	1.4
Plan	5	3.0	.6	2.0	1.0
Relation of plan to essay	5	3.3	.6	2.0	1.0
Introduction	5	2.6	.6	2.0	1.2
Order	5	2.7	.4	2.0	1.2
Emphasis	5	2.2	.5	2.0	1.3
Conclusion	5	2.3	.6	2.0	1.3
Vividness of words	10	5.8	.8	4.0	2.3
Figures of speech	5	2.4	.4	2.0	1.0
Vocabulary	5	2.5	.4	2.0	1.1
Sentence Structure	15	8.1	1.0	6.0	3.0
Sentence beginnings	5	2.8	.3	2.0	.8
Economy	15	8.0	1.0	6.0	3.0
Total impression	15	8.0	1.2	6.0	3.4
Total Style-Content	125	67.8		50.0	26.2

*Total scores for mechanics were not permitted to exceed 125.

The table shows that English mechanics contributed much more than expected. It was surprising to note that spelling, alone, contributed more than 16 'style-content' variables combined. This was definitely not the intention of the curriculum committees in charge of Grade XII English, therefore it must be concluded that a re-examination of the method and weighting of the variables is in order. A re-examination of the actual variables themselves is also indicated, as a result of the factor analysis reported earlier.

A study involving fewer variables, and employing a different weighting technique, in the scoring of essays is in progress.

REFERENCES

CAST, B.M.D. "The Efficiency of Different Methods of Marking English Composition", British Journal of Educational Psychology, 9-10: 257-269, 49-60 (1939 and 1940).

COFFMAN, William E. and KURTMAN, David. "A Comparison of Two Methods of Reading Essay Examinations". American Educational Research Journal, 5: 99-107, 1968.

COWARD, Ann F. "A Comparison of Two Methods of Grading English Compositions". Journal of Educational Research, Vol. 46, No. 2: 1952, p. 81-93.

DARSIE, Marvin L. "The Reliability of Judgments Based on the Willing Composition Scale". Journal of Educational Research, 5: 89-90, 1922.

DIEDERICH, Paul B., FRENCH, John W. and CARLETON, Sydell T. Factors in Judgments of Writing Ability. Princeton, N.J.: Educational Testing Service, 1961.

HUDDLESTON, Edith M. "Measurement of Writing Ability at the College Entrance Level: Objective vs. Subjective Testing Techniques". Journal of Experimental Education, 22: 165-213, 1954.

HULTEN, C.E. "The Personal Element in Teachers' Marks". Journal of Educational Research, 12: 49-55, 1925.

MORRISON, R.L. and VERNON, P.E. "A New Method of Marking English Compositions". British Journal of Educational Psychology, 11: 109-119, 1941.

REMONDINO, C. "A Factorial Analysis of the Evaluation of Scholastic Compositions in the Mother Tongue". British Journal of Psychology, 29: 242-251, 1959.

STARCH, Daniel and ELLIOT, Edward C. "Reliability of the Grading of High School Work in English". School Review, 20: 442-457, 1912.

WINER, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Co., 1962.